Leveraging Convex Hulls for Advanced Statistical Monitoring

Polychronis Economou

Department of Civil Engineering, University of Patras

2024

UNIVERSITY OF PATRAS
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

Ε Ε L
Environmental Engineering Laboratory

Outline

1. Introduction

2. Monitoring the number and the spatial distribution of events on a plane
   - Motivation
   - Monitoring method
   - Application

3. Non-Parametric monitoring of multivariate data stream
   - The problem
   - Monitoring Procedure
   - Simulations

Introduction

Multivariate Statistical Process Monitoring is one of the most rapidly developing areas of statistical process control

Control charts can be very useful in various domains like biosurveillance, data stream analysis, and manufacturing.

Introduction

This study introduces two novel monitoring procedures by integrating **convex hulls** and **chi-square-like statistics** with control charting techniques.

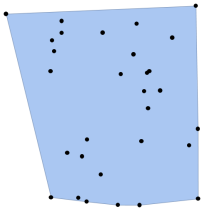
1. monitoring simultaneously the number and the spatial distribution of events on a plane under the assumption of a homogeneous Poisson process

2. a non-parametric monitoring process that aims to detect changes in multivariate data streams

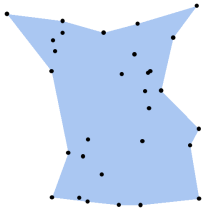
Convex Hull

Definition

A set C is **convex** if the line segment between any two points in C lies in C



Convex



Non-convex

Definition

The **convex hull** of a shape (set of points) is defined as the smallest convex set that contains it (containing all the points).

Chi-square-like statistics

A chi-square statistic is a test that measures how a model compares to actual observed data.

General form

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$\hookrightarrow O$ represents the observed quantity (in the classical approach the frequency for each category)
$\hookrightarrow E$ represents the expected value of the quantity under the assumed model.

Monitoring the number and the spatial distribution of events on a plane

Motivating Examples

Public health threats from Infectious disease outbreaks

↪ Early detection of an outbreak of these diseases or the identification of an area with a high concentration of incidents is crucial for limiting the spread of public health threats

Emergency calls

↪ Identify emergency events or areas with a high concentration of incidents

Environmental monitoring (Chemical or biological incidents)

↪ The number and location of incidents can be useful in locating the start (or the hotspot) of the incidents

Monitoring – Assumption

Monitor simultaneously

- the number and
- the spatial distribution of the events

under the hypothesis that the distribution of events is consistent with a **homogeneous Poisson process**.

Homogeneous Poisson process

A spatial point process N is a homogeneous Poisson process with intensity λ if the two following conditions hold

(i) the number of events in any bounded region E follows a Poisson distribution with mean $\lambda|E|$ where $|E|$ denotes the d-dimensional volume of E, and

(ii) given that there are n events in E, those events form an independent random sample from a uniform distribution on E.

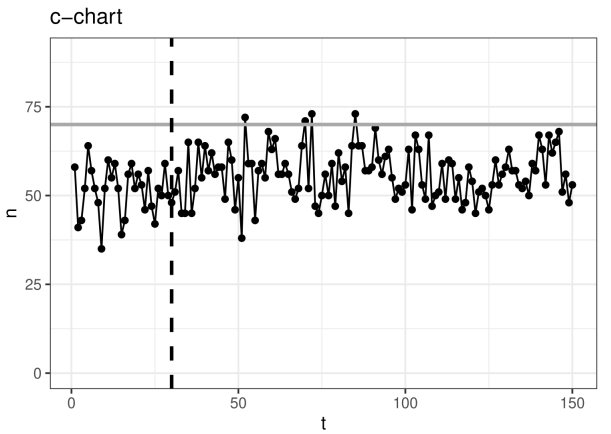
Monitoring Procedure

The proposed method exploits

- a **c-chart** to detect changes in the number of events in an area
- a **convex hull-based chart** to detect any violation of the homogeneous Poisson process assumption, i.e. different concentrations of events into different subareas

Monitoring the number of events

One-sided c-chart for monitoring the mean number of events.



UCL: the $(1-\alpha)$th quantile of the Poisson(λ) distribution for a given α

Monitoring the spatial distribution of the events

Under the homogeneous Poisson process (with known $\lambda = \lambda_0$), the observed sample can be viewed as an independent random sample from a uniform distribution.

So, a test statistic that arises naturally for testing this assumption is

$$D = \sum_{j=1}^{n-3} \frac{\left(1 - \lambda_0 \frac{H_{n-j+1} - H_{n-j}}{H_n}\right)^2}{\lambda_0 \frac{H_{n-j+1} - H_{n-j}}{H_n}},$$

Note: Remove the most remote point to make the test more sensitive

▸ Examples

Monitoring the spatial distribution of the events

Distribution and critical values determination

1. Simulate 10^6 samples under the homogeneous Poisson process assumption by using $n = 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 100, 120, 140, 160, 180, 200$ events on the plane
2. Fit several distributions for each n
3. Identify the best distribution with the smallest AIC
   $\hookrightarrow$ Box-Cox t distribution
3. Use fractional polynomial to describe between the number of events n and each one of the Box-Cox t distribution parameters

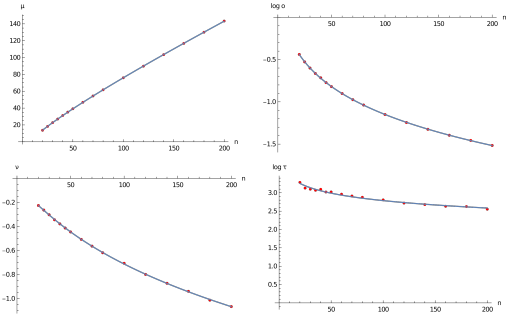
Box-Cox t distribution

$Y \sim BCT(\mu, \sigma, \nu, \tau)$ if the transformed random variable Z

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right] & \text{if } \nu \neq 0; \\ \frac{1}{\sigma} log\left(\frac{Y}{\mu}\right) & \text{if } \nu = 0. \end{cases}$$

follows a (truncated) t distribution with τ degrees of freedom where $\tau > 0$, is treated as a continuous parameter.
The four parameters μ, σ, ν, τ may be interpreted as relating to location and more specifically to the median, scale, skewness, and kurtosis of the distribution, respectively.

Box-Cox t distribution parameters



$$\mu_n = -13.6409 + 3.82936\sqrt{n} + 0.513225n$$
$$\log \sigma_n = 0.348362 - 0.147166 \log n - 0.0386656 \log^2 n$$
$$\nu_n = -0.574495 + 0.389 \log n - 0.0909723 \log^2 n$$
$$\log \tau_n = 4.13109 - 0.291861 \log n$$

Monitoring Algorithm

Step. 1 At each time point record the number n of events and the coordinates of each event.

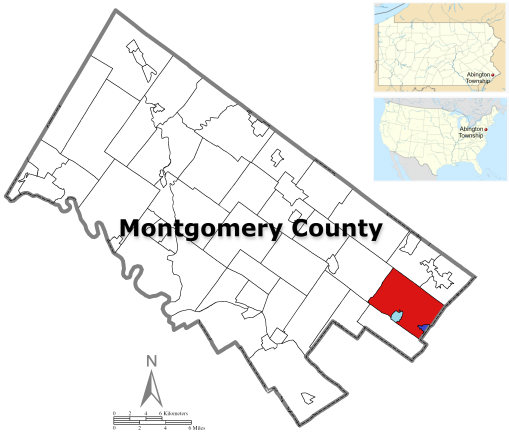
Step. 2 Construct an one-sided c-chart for monitoring the mean number of events with $\alpha = 0.0027$.

Step. 3 Calculate the observed value D_{obs} of the test statistic D, determine the upper 0.0027 quartile under the null hypothesis for the observed number n of events and construct the convex-chart.

Step. 4 Using both procedures (c-chart and convex-chart) determine if the process is under control or not.

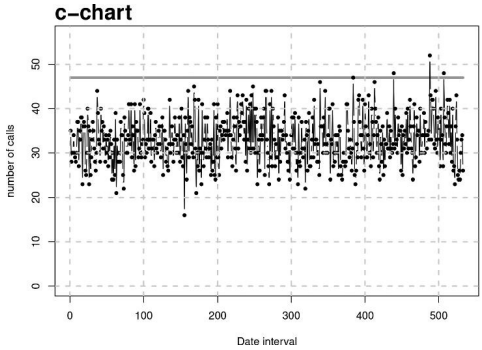
Application

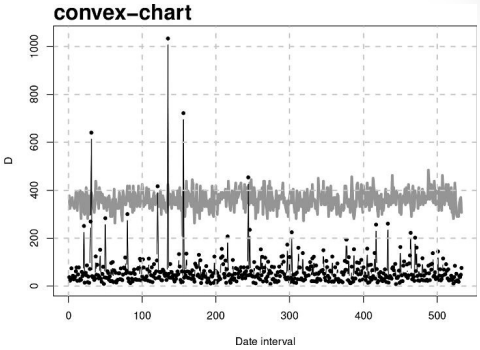
Emergency Medical Services calls to 911, from 12/10/2015 to 4/28/2020, in Abington Township (red), Rockledge (light blue), and Jenkintown (dark blue) in Montgomery County, Pennsylvania, United States



Time interval of three days as a sampling unit, resulting in 533 total samples 17,516 unique records/calls (approximately 10 EMS calls per day)

Application



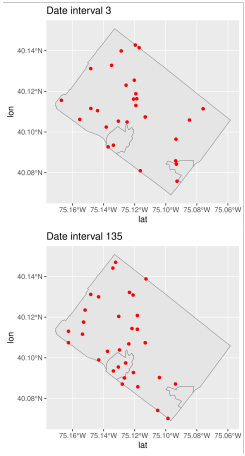


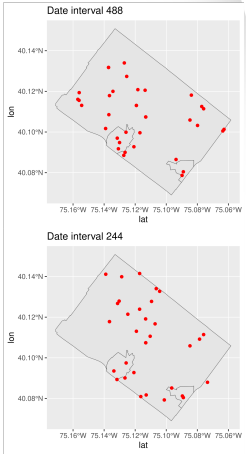
Application

Upper left map
(3^{rd} time interval)
no violation

Upper right map (488^{th})
A large number of calls

Bottom maps
(135^{th} and 244^{th})
a violation of the homogeneous Poisson process assumption was identified





Non-Parametric monitoring of multivariate data streams

Monitoring data stream

The problem

Monitor multivariate data stream
$S_1, S_2, \ldots$

Identify any change in the underlying distribution

The solution

nonparametric, geometric-centric monitoring method based on the convex hull

The idea

If all samples are drawn by the **same distribution** their geometric characteristic should be more or less **similar**.

On the contrary, if **a change occurs** in the underlying distribution this will more likely **be reflected** in the geometric characteristic of the samples.

Monitoring Procedure

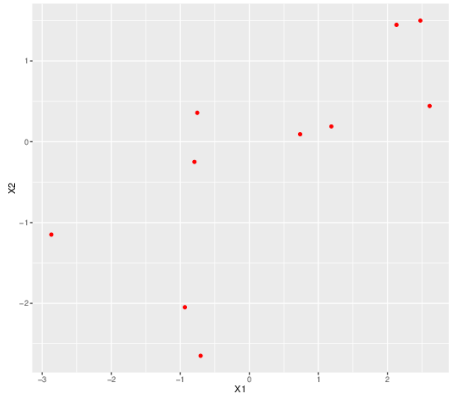
The proposed monitoring scheme relies on

1. **a Phase I analysis** to reveal the unknown geometric structure of the underlying population
2. **a data augmentation retrospective Phase** to estimate control limits for the proposed chart
3. **a Phase II** real-time monitoring prospective procedure

Phase I

Let $S_1, S_2, \ldots, S_k$ be a sequence of d-dimensional samples from Phase I each of size n (relative small sample)

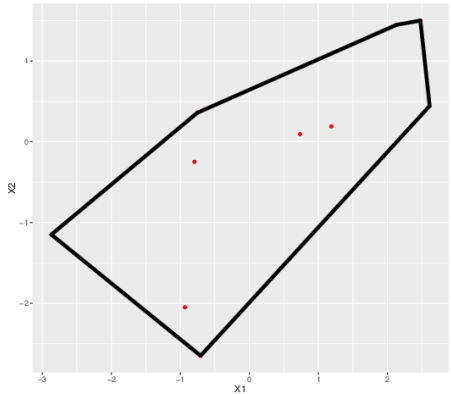
1st sample



Phase I

Let $S_1, S_2, \ldots, S_k$ be a sequence of d-dimensional samples from Phase I each of size n (relative small sample)

1st sample



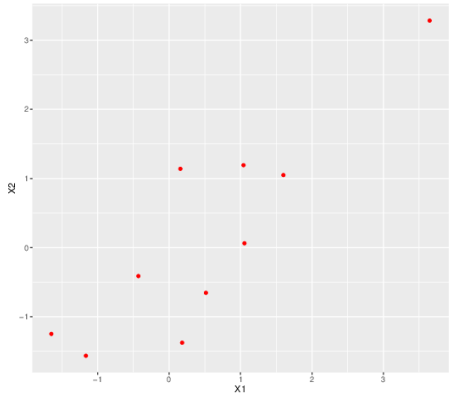
Calculate and store V_{11} of $CH1_1$ Convex Hull.

V_{11}: the length/area/volume or hypervolume (depending on d)

Phase I

Let $S_1, S_2, \ldots, S_k$ be a sequence of d-dimensional samples from Phase I each of size n (relative small sample)

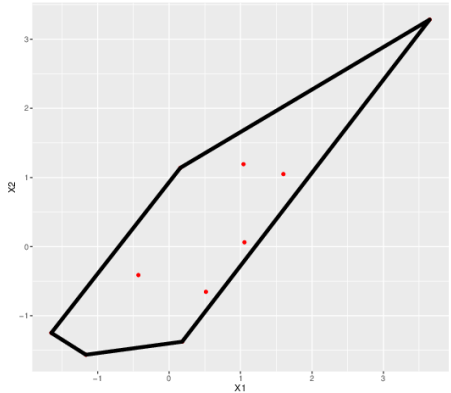
2nd sample



Phase I

Let $S_1, S_2, \ldots, S_k$ be a sequence of d-dimensional samples from Phase I each of size n (relative small sample)

2nd sample

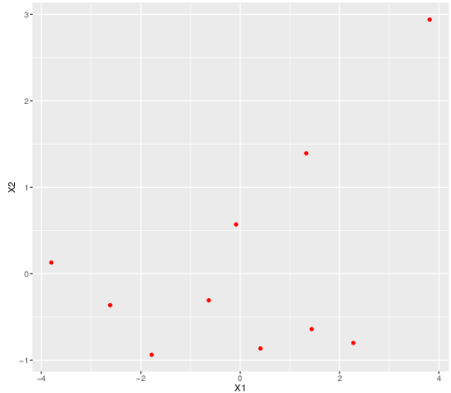


Calculate and store V_{12} of $CH1_2$ Convex Hull.

Phase I

Let $S_1, S_2, \ldots, S_k$ be a sequence of d-dimensional samples from Phase I each of size n (relative small sample)

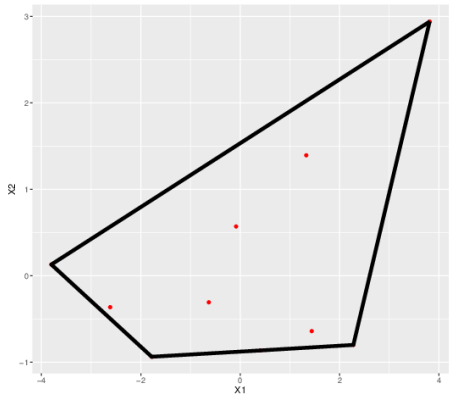
kth sample



Phase I

Let $S_1, S_2, \ldots, S_k$ be a sequence of d-dimensional samples from Phase I each of size n (relative small sample)

kth sample



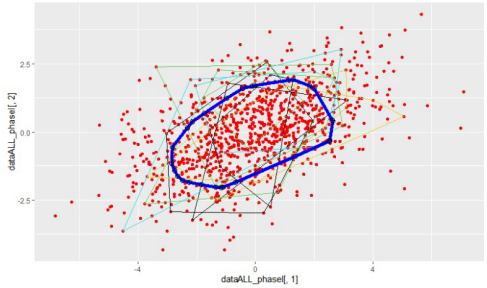
Calculate and store V_{1k} of $CH1_k$ Convex Hull.

Data augmentation Phase

After the k^{th} sample
$\hookrightarrow$ generate k_b bootstrap samples of size n from the pooled data of Phase I are generated
$\hookrightarrow$ calculate and store the volumes $V_{1j}, j = (k+1), \ldots, (k+k_b)$ of the corresponding Convex Hulls.

Determine the **"typical" Convex Hull**, $CH0$

$\hookrightarrow$ Remove the furthest point until the volume of the Convex Hull gets smaller than (or equal to) the **median** volume of the $k + k_b$ Convex Hull volumes.



Data augmentation Phase

Next, the following quantities are calculated

1. $V_{1j}^{*}, j = 1, \ldots, k + k_b$: the volumes of the Convex Hulls, $CH1_j^{*}$, defined by the intersection of each $CH1_j$ after removing the furthest point and $CH0$
2. $V_{2j}, j = 1, \ldots, k + k_b$: the volumes of the Convex Hulls, $CH2_j$, defined by the intersection of each $CH1_j$ and $CH0$
3. $V_{3j}, j = 1, \ldots, k + k_b$: the volumes of the union of $CH1_j$ and $CH0$
4. $V_{4j}, j = 1, \ldots, k + k_b$: the volumes of the Convex Hulls, $CH4_j$, defined of the union of the points of $CH1_j$ and $CH0$

Phase II

One-side control charts are constructed based on the four following test statistics using the corresponding upper α quantiles (obtained by the previous Phase).

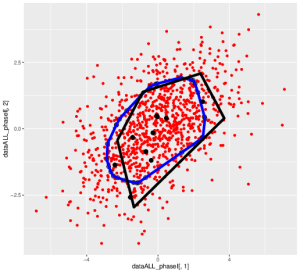
Test statistics

- $\chi_1^2 = \frac{\left(V_{1i} - \bar{V}_1\right)^2}{\bar{V}_1}$
- $\chi_1^{*2} = \frac{\left(V_{1i}^* - \bar{V}_1^*\right)^2}{\bar{V}_1^*}$
- $\chi_2^2 = \frac{\left(V_{2i} - \bar{V}_2\right)^2}{\bar{V}_2}$
- $\chi_{34}^2 = \frac{\left((V_{3i} - V_{4i}) - \left(\bar{V}_3 - \bar{V}_4\right)\right)^2}{\bar{V}_3 - \bar{V}_4}$

Phase II

$$\chi_1^2 = \frac{\left(V_{1i} - \bar{V}_1\right)^2}{\bar{V}_1}$$

$$\chi_1^{*2} = \frac{\left(V_{1i}^* - \bar{V}_1^*\right)^2}{\bar{V}_1^*}$$

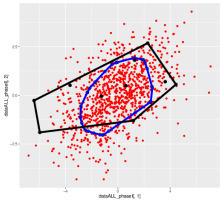


$$\chi_2^2 = \frac{\left(V_{2i} - \bar{V}_2\right)^2}{\bar{V}_2}$$

$$\chi_{34}^2 = \frac{\left((V_{3i} - V_{4i}) - \left(\bar{V}_3 - \bar{V}_4\right)\right)^2}{\bar{V}_3 - \bar{V}_4}$$

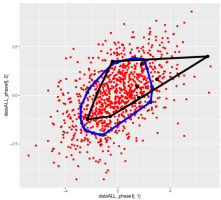
variance increase

+correlation decrease



χ_1^2-chart & χ_1^{*2}-chart

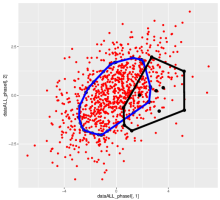
outlier



χ_1^2-chart

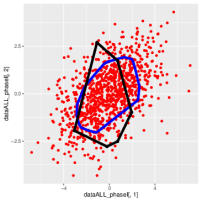
mean shift

+correlation increase



χ_2^2-chart

rotation



χ_{34}^2-chart

Simulation - Set up

Number of simulations=1,000

Phase I: 100 samples, n=10 (in total 1,000 Phase I observations)

Bootstrap samples from the pooled sample of Phase I: 2,000

Distribution: $(X_1, X_2) \sim N_2(\mu, \Sigma)$, where

$$\mu = (0, 0) \text{ and } \Sigma = \begin{bmatrix} 4 & 1.414 \\ 1.414 & 2 \end{bmatrix}$$

Simulation – Scenarios

- In control
- Out of control
  1. Variance increase of X_1 (from $Var(X_1) = 4$ to 6 and 8)
  2. Single outlier (add 1, 2 or 3 standard deviations in X_1 to the observation with the larger x_1)
  3. Shift in the mean of X_1 by 1, 2 and 3 standard deviations
  4. Correlation change (from 0.5 to 0.3, 0.4, 0.6 and 0.7)
  5. Rotated data (20°, 40°)

Simulation - Results

		χ_1^2			χ_1^{*2}			χ_2^2			χ_{34}^2		
		ARL		SDRL	ARL		SDRL	ARL		SDRL	ARL		SDRL
In-control		388.431		485.303	366.850		456.400	347.717		393.749	369.455		488.991
			$\Delta\%$			$\Delta\%$			$\Delta\%$			$\Delta\%$	
Scenario 1	$4 \to 6^*$	35.048	**90.98**	38.784	39.849	**89.13**	43.189	495.444	-	560.866	128.416	65.24	263.538
$var(X_1)$ increase	$4 \to 8^*$	11.630	**97.00**	12.401	12.119	**96.70**	11.280	511.109	-	563.475	58.527	84.16	191.940
	$^*\approx +20\% - 30\%$ simultaneous alerts from χ_1^2 and χ_1^{*2} compared to the other scenarios												
Scenario 2	1 s.d.	74.161	**74.40**	99.443	121.208	66.96	145.338	869.580	-	941.233	111.596	69.80	195.777
Single outlier	2 s.d.	16.396	**95.78**	18.956	75.160	79.51	84.441	946.012	-	994.197	32.157	91.30	94.685
$maxX_1 \to maxX_1+$	3 s.d.	5.913	**98.48**	6.606	67.958	81.47	78.621	1010.014	-	1046.453	11.226	96.96	41.654
Scenario 3	1 s.d.	388.431	0	485.303	366.850	0	456.400	6.361	**98.17**	8.658	53.714	85.46	95.582
X_1 mean shift	2 s.d.	388.431	0	485.303	366.850	0	456.400	1.084	**99.69**	0.302	2.1670	99.41	2.539
	3 s.d.	388.431	0	485.303	366.850	0	456.400	1.002	**99.72**	0.044	1.016	99.72	0.117
Scenario 4	0.3	144.386	**62.83**	175.789	143.718	60.82	167.883	404.004	-	479.274	215.764	41.60	419.173
Correlation change	0.4	214.958	**44.66**	285.042	204.899	44.15	246.018	380.884	-	437.211	279.458	24.36	407.095
from 0.5 to	0.6	858.157	-	928.214	838.071	-	921.920	265.737	**23.58**	308.037	463.930	-	698.838
	0.7	1708.588	-	1352.042	1630.650	-	1360.407	166.053	**52.24**	231.268	420.251	-	741.702
Scenario 5	20°	388.431	0	485.303	366.850	0	456.400	264.360	23.97	320.246	129.947	**64.83**	305.522
Rotated data	40°	388.431	0	485.303	366.850	0	456.400	140.705	59.53	186.441	24.782	**93.29**	165.528

Simulation - Next steps

↪ Compare with other methods

↪ Use different distributions

Thank you for attention

Email: peconom@upatras.gr
Webpage: http://www.des.upatras.gr/amm/economou/index.htm